Forecasting Traffic Algorithms in Telecommunications

Efstathios Tsolakis

SID: 3308180025

SCHOOL OF SCIENCE & TECHNOLOGY

A thesis submitted for the degree of

Master of Science (MSc) in Data Science

DECEMBER 2019

THESSALONIKI – GREECE

Forecasting Traffic Algorithms in Telecommunications

Efstathios Tsolakis

SID: 3308180025

Supervisor: Prof. Agamemnon Baltagiannis

SCHOOL OF SCIENCE & TECHNOLOGY

A thesis submitted for the degree of

Master of Science (MSc) in Data Science

DECEMBER 2019

THESSALONIKI – GREECE

Abstract

This dissertation investigates time series analysis on Telecommunication Traffic data. The dataset is comprised of two time series of 4G-LTE technology, where one represents the Downlink and the other the Uplink speed for one Base Station. The time series has an hourly frequency and the training set has a total of thirty-seven (37) days and the test set up to sixteen (16) days. We initially perform a Visual and Statistical exploration to gain insight for applying the according algorithms. Data preprocessing was necessary since we have real-life data, where missing and erroneous values occur. Also, a look into statistical and econometric approaches takes place, in order to find the optimal algorithms with respect to the special characteristics of the given data. Briefly, the dissertation describes the procedure for using Python as the tool to efficiently and accurately provide forecasts on a time series, to gain knowledge for effective management and planning for a Telecommunication network. Lastly, the dataset used was for the months February to April of 2014, in Greece.

Efstathios Tsolakis

December 2019

Acknowledgements

I would like to thank my supervisor, Dr. Baltagiannis for his guidance and consulting throughout the work of this dissertation. Also, my fellow students and friends for working productively and enjoyably on team projects during the master's degree. Most of all I would like to thank my family and especially my brother, Angelos, for all the love and support I wish to return.

Contents

1 Introduction

Mobile telecommunication networks play a vital role in everyday life. They are a necessity for a functional society since now people rely on instant communication through their mobile device. The communication could either be in voice or in video, for matters such as work-related interaction, personal matters; communication with family, friends etc. Also, entertainment now passes through mobile devices, for uses in gaming, news, blogs, social media sites, photography, with the majority of data being transferred for viewing videos [1].

In this dissertation, various approaches are followed in forecasting the traffic of a mobile telecommunication network. These approaches will be modelling with econometric methods such as ARIMA, Exponential Smoothing and variations of these. Having created these models, we will attempt to:

- Perform Exploratory Data Analysis and Visualization of the data
- Compare results from individual models
- Compare results on long-term and short-term predictions
- Create a hybrid of the previous models to achieve optimal forecasting
- Propose models for future works

Almost every decade, technology advances in the mobile telecommunication sector are deployed. These technologies are named First Generation, Second Generation up to Fifth Generation, where the latter started deployment in commercial networks in Spring 2019. Every time a new Generation is deployed, they increase the bandwidth speed up to 100 times. To be more specific, the download link speed of 3G is 7.2 Mbps where in 4G it can reach up to 300 Mbps [2] and in 5G technology the maximum download speed that can be achieved is up to 20 Gbps.

There is a significant importance for both mobile phone users and Mobile Network Operators -that provide cellular service to the users, to understand patters of mobile traffic data. Accurately forecasting the demand for data traffic can be useful for the service

providers to schedule network maintenance of the cell towers. Also, they can offer optimal service regarding Quality of Service indicators. Moreover, allocating large amounts of expensive equipment in areas with low demand and vice versa, lower quality equipment to areas with high demand is poor resource utilization and cost-expensive for the Mobile Network Operator. Simultaneously with minimalization the cost for infrastructure and machinery, accurately forecasting the traffic leads to efficient electrical power management of the cell stations.

The increasing demand in mobile traffic can be seen in [3]**,** where the monthly traffic in 2017 was 12 exabytes and is anticipated to reach 77 exabytes monthly by 2022. Cellular traffic will consist of the basis of revolutionary technologies such as Augmented and/or Virtual Reality, Autonomous Vehicles and Digital Healthcare [4]

2 Mobile Networks

In this chapter, the history of Mobile Networks is investigated. The rapid development shows not only their importance, but also how Data is now embedded within them. As a continuation, various use cases of Data Science in Telecommunication Systems will be presented.

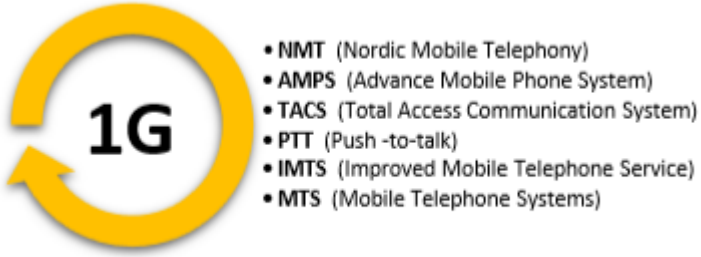
2.1 Generations

The stages of development in Mobile Telecommunication Networks are characterized by Generations. This method of categorization was applied after huge advancements that each can be distinguished.

2.1.1 First Generation

In 1981 the first generation of wireless mobile telecommunication network was deployed. The system is based on analog signals with Frequency Modulation (FM) and frequency division multiple access as the basis of this technology. However analog technology does not provide the transmission of data and security. Also, interference and high battery consumption are other main issues tackled by the Second Generation of Cellular Networks [5]. The standards of 1G are depicted in the figure 1.

Figure 1 [5]



2.1.2 Second Generation (2G)

The implementation of 2G technology took place in 1991. Digital Modulation replaced the frequency modulation in the 1G networks, offering services as voice and data transmission. The first 2G network is known as GSM and supports data rate transmission from 14.4 kbps up to 384 kbps. Also, with digitalizing the signal, noise interference was drastically reduced, and voice quality and security improved significantly. Different technologies were introduced that all belong to the 2G family, such are GPRS and EDGE [5].

2.1.3 Third Generation (3G)

The first commercial 3G network was launched in Japan, on October 1, 2001. Data bandwidth varies from 128 kbps for fast moving users (e.g. trains, cars, etc.) up to 2 Mbps for fixed wireless LANs. Until 3G technology, different countries used different standards in 2G technology. Therefore, the necessity of a global standard arose with 3G which is a family of Cellular Networking standards that can work together [6] . The International Telecommunication Union (ITU) is the founder of 3G technology and an organization called the 3rd Generation Partnership Project has continued the work to define a mobile system alongside the ITU's standards.

New services introduced to mobile phone users with this advanced are multimedia services, video conferencing, gaming and high-speed internet connectivity [7].

2.1.4 Fourth Generation (4G)

The deployment of 4G technology took place in 2010, after 10 years of research. Speeds up to 1 Gbps are achieved with 4G, 500 times faster than the predecessor. An extension to the 3GPP was achieved in 4G technology, namely All-IP, which establishes a common platform for all previous technologies to harmonize and collaborate. These are GSM – Global System for Mobile Telecommunications, GPRS – General Packet Radio Service, IMT-2000, Wi-Fi and Bluetooth. Less expensive data transfer, enhanced roam-

ing capabilities and unified messaging and broadband multimedia are the new interactive services followed with 4G technology [2].

The following are some of the standards required for the IMT-Advanced in 4G:

- Download speed up to 1 Gbps and upload speed up to 500 Mbps.
- Latency time for connection and synchronization of 100ms and 10ms respectively.
- Download and Upload spectral efficiency of 15 bps/Hz and 6.75 bps/Hz respectively.

2.1.5 Fifth Generation (5G)

The fifth generation of communication standards was first adopted on a large scale in April 2019 and is currently in progress. Main benefits include download link speeds of up to 20 Gbps when a user is in low mobility, meaning travelling with low speeds, such as walking. Also, the connections latency between a user and the server is smaller than 1ms, ranging from 10 to 100 times lower than in 4G. Also, 5G infrastructure can support an increased number of devices connected and a higher capacity of traffic data within an area (1000 higher than LTE 4G) [1].

Since the number of devices capable to connect via the antennas in a 5G network increases drastically, the opportunity is given for a highly connective Internet of Things infrastructure. Billions of new devices will be connected via Wi-Fi or 5G, according to the authors in [8]. It is estimated from a report by Ericsson that by the year 2022, 1.5 billion cellular devices will be connected to the network. In 2019 the total number of connected IoT devices reaches 20 billion and by 2022 will total of 30 billion [9].

Therefore, 5G technology enhances user experience in existing applications, such as education, entertainment, manufacturing and many others. Also, 5G comprises a gateway to multiple revolutionary technologies such as Autonomous Driving and Smart City Grids. A table is shown from [10] describing these emerging technologies.

Table 1 [10]

Verticals	Drivers	Enablers	5G requirement
Education	• Remote delivery • Immersive experiences	• Video streaming • Augmented reality/ • Virtual reality	• Large bandwidth • Low latency
Manufacturing	• Industrial automation	• Massive IoT networks	• High connection density • Ultra reliability • Low power consumption
Healthcare	• Remote diagnosis and intervention • Long term monitoring	• Video streaming • Augmented reality/ Virtual reality • Embedded devices, advanced robotics	• Low power • High throughput • Low latency
Smart Grid	• Intelligent demand/ supply control • Powerline communication	• IoT sensors and networks	• High reliability • Broad coverage of network • Low latency
Entertainment	• Immersive gaming and media industry • Multimedia experience at 4k, 8K resolution	• Video streaming • Augmented reality/Virtual reality	• Large bandwidth • Low latency

Below we show a table comparing generations from 1G up to 5G.

Generation	1G	2G	2.5G	3G	3.5G	4G	5G
Start	1970-1980	1990-2000	2001-2004	2004-2005	2006-2010	2011-Now	Soon (2020)
Data Bandwidth	2 Kbps	64 Kbps	144 Kbps	2 Mbps	More than 2 Mbps	1 Gbps	more than 1 Gbps
Technology	Analog Cellular	Digital Cellular	GPRS, EDGE, CDMA	CDMA 2000 (1xRT, EVDO) UMTS, EDGE	EDGE. Wi-Fi	WiMax LTE Wi-Fi	wwww
Service	Voice	Digital Voice, SMS,Higher Capacity Packet Size Data	SMS, MMS	Integrated High Quality Audio, Video & Data	Integrated High Quality Audio, Video & Data	Dynamic Information access, Wearable Devices	Dynamic Information access, Wearable Devices with AI Capabilities
Multiplexing	FDMA	TDMA, CDMA	CDMA	CDMA	CDMA	CDMA	CDMA
Switching	Circuit	Circuit, Packet	Packet	Packet	All Packet	All Packet	All Packet
Core Network	PSTN	PSTN	PSTN	Packet N/W	Internet	Internet	Internet
Handoff	Horizontal	Horizontal	Horizontal	Horizontal	Horizontal	Horizontal & Vertical	Horizontal & Vertical

Figure 2 [11]

2.2 Radio Access Networks

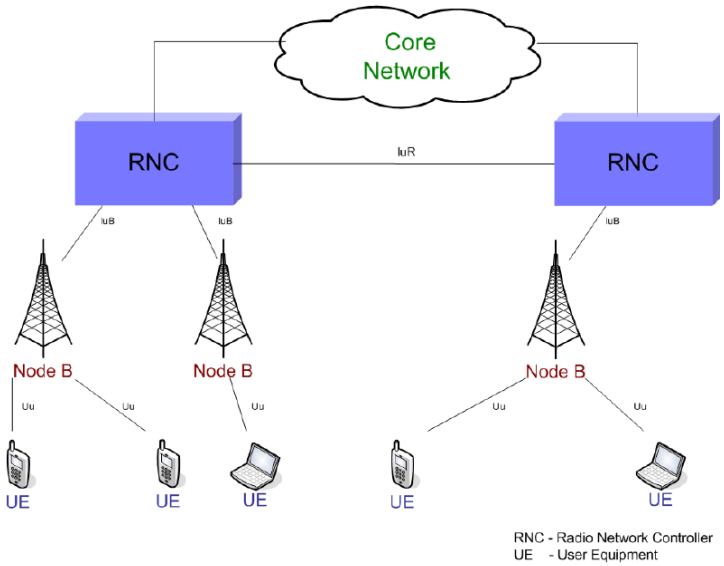
The connection of individual devices to a telecommunication network is achieved through radio connections. Therefore, such a Telecommunications System is named a Radio Access Network (RAN). A RAN is how an end-user with a mobile device is connected to a core network, such as the internet. Also, it is considered a principal component in Telecommunication Systems. The coordination and management of connections of end-users to the core network is achieved by the RAN. Also, a mobile device may be connected to multiple RANs, therefore, handling these connections is a fundamental process within the RAN.

The basic components of a RAN are a Base Station (BS), with the according antennas, and the Radio Network Controller (RNC). In 4G-LTE technology, the Base Stations are referred to as eNodeB, or Nodes [12], [13].

2.2.1 RAN Controller

The main process of the RAN controller is to control the nodes connected to it. This is done with subprocesses as resource management, mobility management and data encryption.

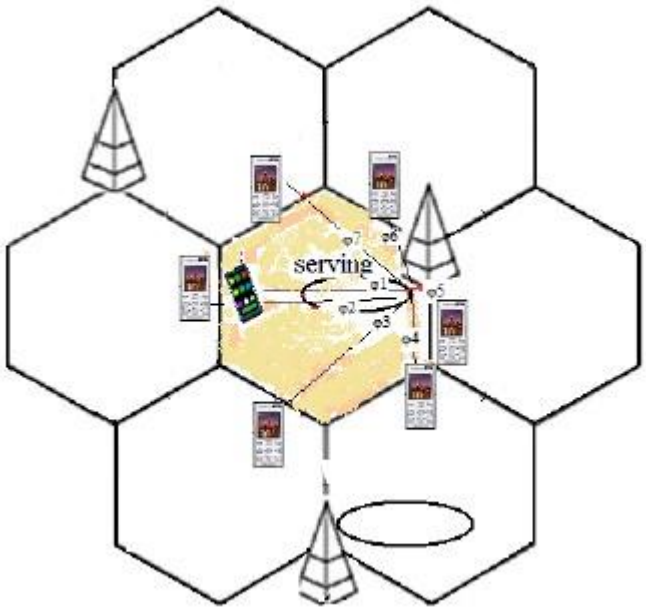
Figure 2 [13]



2.2.2 RAN base station

The base station resides between the end-user and the RAN controller and are responsible for taking digital packets from the core network and transmitting them with radio signals to the end-user device. A Node size differs from 1 sector up to 3 sectors, depending on the range and capacity of the area they wish to cover. Sectors are the antennas used for transmission and cover a circle arc with degrees varying from 60 to 120. The authors in [14] note that in LTE networks, the responsibility of a Base Station is the coverage area, the calculation of transmission power.

Figure 3 [14]



2.3 Data Science in Telecommunications

Throughout the boom of technological advances in the Telecommunications sector, enormous amounts of data are being generated. Alongside this process, the means and methods for processing this data is also rapidly developing with Data Science. The high business insight and value that Data Science can generate, is considered a necessity in the Telecommunication Business. There are a wide variety of applications that are used to maximize company profits, minimize costs, improve marketing and business strategies and in summary to process, visualize and extract critical components for efficient business solutions.

Here, general categories are presented of how Data Science can be harnessed into the Telecommunications industry [15].

2.3.1 Fraud Detection:

Fraud is the intentional act of deception in order to gain unfair or unlawful gain against a victim. Consumers are exposing sensitive personal information with the use of telecommunications, for applications such as e-banking and performing business. This generates a potential for malicious users to exploit.

The authors in [16] review how an American company, namely AT&T, systematically attempted to address major fraud schemes.

A few examples are given on some common varieties of fraud in telecommunications. These are:

- Subscription fraud: Signing up for a service without paying the according fees
- Intrusion Fraud: Illegitimately selling calls to third parties based on an account
- Fraud based on loopholes: Exploiting passwords for mailbox services and proceeding to make outgoing calls.
- Social Engineering: Impersonating a customer to access sensitive information from the telecommunication company
- VOIP: rerouting their phone number to a service of a different company, usually in another country, for lower prices

2.3.2 Customer Churning:

Customer churn is an evaluation metric for expressing the percentage of customers who cease using a company's services and switch to a competitor. Another widely used term for expressing churning is customer attrition and is a subject in Customer Attrition Analysis. A high value of churn prediction is due to a smaller cost to maintain a clientele then to invest in new marketing strategies and campaigns to gain new customers.

The authors in [17] present features for inserting into Machine Learning models for classifying a customer as a potential churn. These features include:

- Demographic profiles: age, social class, gender.
- Customer account information: payment types, initial date of subscription, number of calls.
- Telephone line information: services such as voice mail, number of telephone lines, line types.
- Complaint Information: whether complaints have been made and on which dates.
- Historical Information on bills and payments: call details such as total duration, call fees, calls on landlines or mobiles, international calls
- Call details: analytical information regarding each call and not aggregates from the previous
- Incoming call details: duration, number of received calls

The Machine Learning classifiers used in this approach include:

- Artificial Neural Networks
- Support Vector Machines
- Decision Trees
- Naïve Bayes Classifier
- Logistic Regressions
- Linear Classifiers

2.3.3 Network Management:

This discipline regards to evaluating, maintaining and improving the quality of the telecommunication network. The authors in [18] separate use cases of big data into 3 layers: Customer, Service and Resource layers.

In the Resource layer, belong Capacity Management and Network Planning. In the resource layer, tasks such as monitoring network performance of the Cellular Base Stations, routers, switches and other devices takes place to be able to verify and enforce smooth operation of the system. Furthermore, the data processed involves Key Performance Indicators (KPI) of traffic load, latency and packet loss ratio. These KPIs can further be used for network planning, fault management, capacity management with the value generated from Data Science techniques for accurate prediction and Forecasting.

2.3.4 Customer Satisfaction:

A new measure is created in the Telecommunications industry, namely QoE. Quality of Experience differs from the well-known Quality of Service, as the latter measures the quality of network services and connectivity. These are parameters that can be measured quantitively and therefore be expressed in definitive numbers. Some of the most popular are packet loss, latency, jitter, bit rate, etc. In contrast, Quality of Experiences is a subjective measure of a customer's satisfaction in a scale from 1 to 10. There are some quantitative measures that can be inserted, such as call duration, voice & video data, number of calls, however these remain subjective per user. Benefitting from measuring the Customer Satisfaction (CS) can be seen in other use cases of Data Science in Telecommunications. These include predicting customer churning and target marketing. An approach for quantifying CS can be established by using quantitative variables -as mentioned above, and service KPIs. Sessions are a time frame of the connection between a user communication with an Internet service, therefore the KPIs can be split into sessions called S-KPIs for real-time analysis. The authors [19] insert user feedback as objective metrics may not align with a user's subjective experience.

3 Theory and Previous Works

One of the main algorithms used in time series forecasting is based on Autoregressive and Moving Average processes. With combining these processes, ARMA models are created, with plenty of varieties being created upon these models. Theory on how these models work will be covered. Another algorithm widely used for forecasting time series, namely Exponential Smoothing, will be investigated also.

Also, in this Chapter an investigation on previous research applied onto Telecommunication data takes place.

3.1 Arima models

The Autoregressive Integrated Moving Average (p,d,q) model is a statistical approach to understand and forecast a time series. Two main parts comprise an Arima model, an Autoregressive process and a Moving Average process. In the AR process, the weighted average of past values forecast the next value. On the other hand, the Moving Average process defines the next value from the previous residual errors.

3.1.1 Stationarity and the Backshift Operator

Stationarity is an important factor in Arima time series analysis, that requires the mean and the variance to be constant over time. If a time series is not stationary, it may be having a trend or a seasonal factor. For this to be achieved, Differencing Operators are applied to the values of the time series and this process is express by the term Integrated. Depending on the lag values necessary to be differenced, the parameter d is found. The simplest form is of value 1, named the first differencing operator and is defined as:

$$\nabla x_i = x_i - x_{i-1}$$

3.1.2 AR Process

The AR process is the backbone of complex models. The simplest form of an Auto-regressive (AR) model is of order 1. The equation of such a model takes the form:

$$y_t = \varphi_0 + \varphi_1 y_{t-1} + \alpha_t$$

where y_t is the value at time t, φ are constants for the lags that have a value that ranges from -1 to 1 that are determined through the model and α_t is white noise, also named as an error term.

3.1.3 MA Process

The Moving Average process is of equal importance with the Autoregressive. The simplest form of first order is shown below:

$$y_t = \varphi_0 + \varepsilon_t + \theta_1 \varepsilon_{t-1}$$

Where ε is a white noise time series with a mean of zero and a variance of σ^2. The parameter θ takes values from -1 to 1.

The difference with the Autoregressive process is that in AR models, the lag term is multiplied by a constant to predict the next value y_t, where in the Moving Average process, the MA terms are the past errors.

3.1.4 ARMA Process

Here the process is a combination of the two previous into one model. With only two parameters, an ARMA(1,1) process takes the form:

$$y_t - \varphi_1 y_{t-1} = \varphi_0 + \varepsilon_t + \theta_1 \varepsilon_{t-1}$$

where α_t is a white noise series. The left side of this equation is the Autoregressive component where the right side the Moving Average.

The general form of an ARMA(p,d) process takes the form [20]:

$$y_t = \sum_{i=1}^{p} \varphi_i y_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}$$

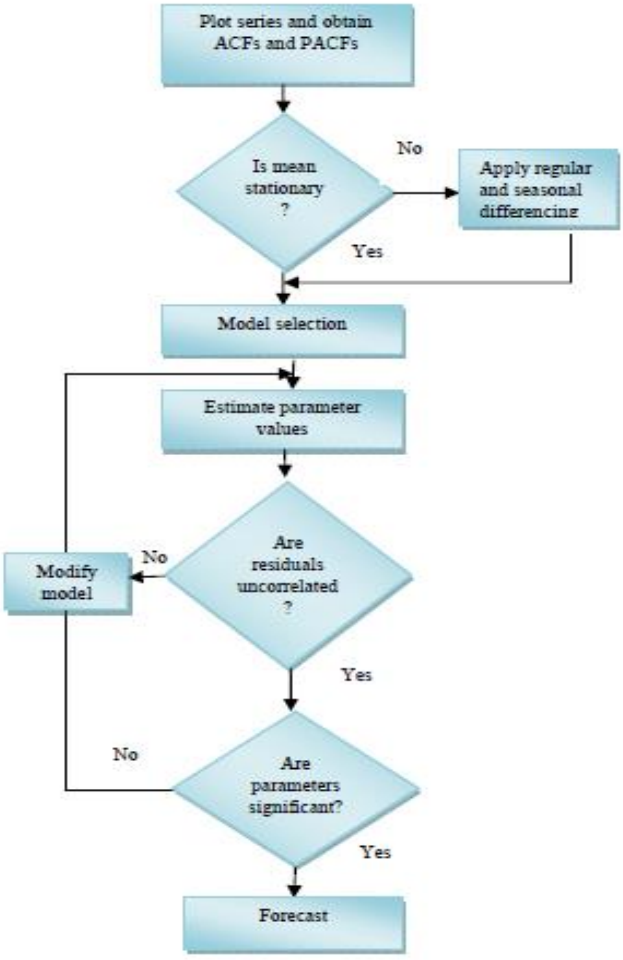
3.1.5 ARIMA

The ARIMA model is an evolution of the ARMA process, where the full name is Auto-regressive Integrated Moving Average. The Integration term applies differencing on the terms y_t with a previous one and is expressed with the use of the Backshift Operator.

A good measure of the series' stationarity is given by the autocorrelations function (ACF). If a gradual decay is found within the significant lags of the ACF, there is a strong indication of correlation among the observations at different times.

The process of finding the statistically important parameters p, d, q is shown in the flowchart below:

Figure 4 [21]



Non-seasonal Arima

The final form of a non-seasonal ARIMA model takes the following form:

$$y'_t = c + \varphi_1 y'_{t-1} + \cdots + \varphi_p y'_{t-p} + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t$$

Where y'_t is the differenced series, of one or multiple times, ε_τ is the random error with zero mean and a constant variance of σ^2.

Seasonal Arima:

The seasonal Arima model takes consideration of characters of a seasonal nature and applies differencing within a time period of the repeating seasonal pattern. For this to be

achieved, a set of three (3) new parameters are implemented and a seasonal ARIMA model is expressed as SARIMA (p,d,q) (P,D,Q)S, where:

- p = non-seasonal AR order
- d = non-seasonal differencing
- q = non-seasonal MA order
- P = Seasonal AR order
- D = Seasonal differencing
- Q = Seasonal MA order
- S = Period of the season.

Mathematically the SARIMA (1,1,1) x (1,1,1)S model is expressed as:

$$(1-\varphi_1 B)(1-\Phi_1 B^S)(1-B)(1-B^S)y_t = (1-\theta_1 B)(1-\Theta_1 B^S)\varepsilon_t$$

Where B represents the backshift operator such that:

$$By_t = y_{t-1}$$

3.2 Exponential Smoothing

In opposition to Arima models, Exponential Smoothing (ES) is based on simple mathematical procedures that assign weights to past observations in a time series. The weights decay with a specific rate, named the smoothing factor. Smoothing factors are usually defined as *alpha* and take values from 0 to 1. Therefore, a low smoothing factor with a value close to 0 allows old observations to have higher influence on the forecast than a smoothing factor with a value close to 1. This method, ES, is used mostly for short-term forecasting, as high it is more reliable for short time windows.

Built upon this idea, some varieties of ES can be constructed.

3.2.1 Simple Exponential Smoothing

The foundation of Exponential Smoothing is built upon SES. Here we have only one (1) smoothing factor *alpha*. The mathematical form is:

$$F_{t+1} = aA_t + (1-a)F_t$$

Where F_t is the forecast at time t and A_t is the actual value.

The initial value for the forecast F_2 at the second period cannot be calculated in this recursive manner. Therefore, we set a constraint that:

$$F_2 = A_1$$

3.2.2 Double Exponential Smoothing

By adding a new smoothing factor, namely gamma, γ, we can capture trends within a time series. Also, this provides a new equation to the procedure:

$$b_{t+1} = \gamma(S_{t+1} - S_t) + (1-\gamma)b_t$$

So now with the new equation for the series b, we update the initial ES equation to:

$$F_{t+1} = aA_t + (1-a)(F_t + b_t)$$

[22].

We notice that Double Exponential Smoothing is a recursive application of the Simple Exponential Smoothing. Many other forms have been proposed, such as Triple Exponential Smoothing, Quadruple and so on, depending on the depth of the recursion.

Other variates of ES depend on the initial values that are given to the model. Since for the forecast at time period 1, we cannot investigate past observations to calculate it, the initial values could be of the form:

$$b_1 = y_2 - y_1$$

$$b_1 = \frac{1}{3}[(y_2 - y_1) + (y_3 - y_2) + (y_2 - y_1)]$$

$$b_1 = \frac{y_n - y_1}{n - 1}$$

The previous equations describe Exponential Smoothing of an additive model, meaning that the trend is linear. Other variations used to capture exponential trends are called *multiplicative*, where the forecasts are not added but multiplied.

3.3 Model Scoring and Metrics

In order to determine the optimal model, we use the parsimony principle. This principle gives a trade-off between a model's accuracy and its complexity. To be more specific, the accuracy can be determined by the Root Mean-Squared Error (RMSE), namely the square of the distance between the predicted and the actual value. In opposition to accuracy, a model's complexity is determined by the number of parameters found within a model. In the literature, two main metrics are used. These are the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC).
The Akaike Information Criterion is defined as:

$$AIC = \mathrm{n}\ln(^{RSS}/_{n}) + 2K$$

where n is the number of data points, RSS is the Residual Sum of Squares and K the models number of parameters, or orders. The RSS is divided by n giving the Root Mean Squared Error [23].

The Bayesian Information Criterion is defined as:

$$BIC = kln(n) - \chi^2$$

where χ2 is the deviance and k the number of parameters in the model [24].

3.4 Previous Works

The authors in [25] perform time series analysis to telecommunication traffic from the country Ghana. Specifically, the traffic under analysis is 2G voice traffic with 700 daily observations. First, they examine the time series for structural breaks, which are distinct changes of the pattern. These sudden changes carry an error to the next observations to be predicted, therefore it is necessary to fix these values. Having fixed the structural breaks, they continue to create a non-seasonal ARIMA (p, d, q) model. For these values to be determined, further analysis of Autocorrelation and Partial Autocorrelation for the lags is applied. Later, they examine the stationarity of the time series with the Zivot-Andrew test and it is found to be non-stationary. To fix this they difference the daily values with the backwards shift operator, given that the series is non-seasonal. By making the time series stationary, they proceed to find the Lag values to create an ARIMA model, as done previously.

They create 6 different ARIMA models, (1,1,1), (3,1,2), (2,1,1), (1,1,2) and (2,1,2) and proceed to testing them in order to find the most statistically significant set of parameters. They conclude to ARIMA (1,1,1).

In [26], the authors use data provided by Ethio Telecom for the time period of October 2015 until June 2016. As Ethio Telecom is the only MNO in Ethiopia, the data shows the total demand of mobile traffic data for this period. In their review of the literature, they state that low order ARIMA models or complex Artificial Neural Network models are suitable for capturing long range dependencies, however for short dependencies Artificial Neural Networks show better results than ARIMA. From plotting the data, they observe and increasing trend along with a seasonal factor of the time series. By finding the ACF and PACF, they find an optimal lag of 7 for differencing the time series to detrend and de-seasonalize the data. They continue to calculate KPSS tests, find a P-value of 0.1 which verifies the stationarity of the adjusted time series. Later, a grid search is completed in order to find the optimal values for the non-seasonal and seasonal components of the SARIMA model. By comparing the errors of the grid search, the model with the highest accuracy is found to be SARIMA (2,0,1) x (0,1,1)7, with a Mean Absolute Percentage Error (MAPE) of 1.17%.

The authors in [27] apply Seasonal ARIMA forecasting algorithms. First, they comment on their dataset, which uses the Erlang metric to describe the traffic. Their data is collected hourly, for multiple cell stations and include a record time. They notice the seasonality within their time series and proceed to underline the two seasonal factors. The first is and hourly seasonality with peaks around midday and minimum values during night. The second seasonal factor found in their time series is on a weekly basis, with higher traffic from Monday until Friday and less traffic during weekends. Later, they import their data into a database by transforming it into a database format. Also, they apply preprocessing tasks on their real data, such as fixing missing and erroneous values due to system faults. In the modelling section, they explore whether differencing is necessary within the time series and compute statistics, such as the autocorrelation function to identify the model's parameters. Having found candidate parameters, they compare the models with metrics such as the AIC (Akaike Information Criterion). The model with the least AIC is chosen, where models do not contain parameter values- named orders, higher than 2. The final model chosen was the Seasonal ARIMA (2,0,1) x (2,1,0)168. The final scoring of their model is achieved by calculating the Normalized Root Mean Squared Error (NRMSE) with values in the range of [0.0810, 0.3618] for 246 of the 269 cells they acquired data from. Also, it should be denoted that they forecasted the next 7 days of a dataset with 28 days.

The authors in [28] forecast traffic on Mobile Network Operators Cells. Specifically, they apply Exponential Smoothing on GSM/GPRS mobile networks for short-term predictions in order to give priority and overload warnings to high traffic cells on rush hours (e.g. downtown) and for capacity planning for the appropriate number of transceivers at the Base Stations. First, they explain their choice of Exponential Smoothing over ARIMA models, as the ARIMA models give accurate results, however, they require high level skills of a researcher to systematically identify and evaluate the time series models for them to satisfy plentiful limitations, which leads up to high expenses. However, Holt-Winters Exponential Smoothing can effectively be applied with inexpensive cost. They continue with analyzing the patters within their data and separate three patterns. An hourly, a weekly and an accidental. The accidental pattern is national holidays, accidents and other unpredictable events. These random events are impossible to be forecasted without sufficient and external data. They chose from several cells on

which to perform their analysis and evaluate the ones with medium or high traffic for these are most interesting and profitable with a dataset of 24 values per day for 25 days. After, they split the data into a training set of 18 days and the remaining into a test set. They chose the RMSE as the metric to compare their forecasting methods. Accuracy is defined by the normalized RMSE and have best results of NRMSE=0.08 on the best cell station and NRMSE=0.33 on the worst cell station.

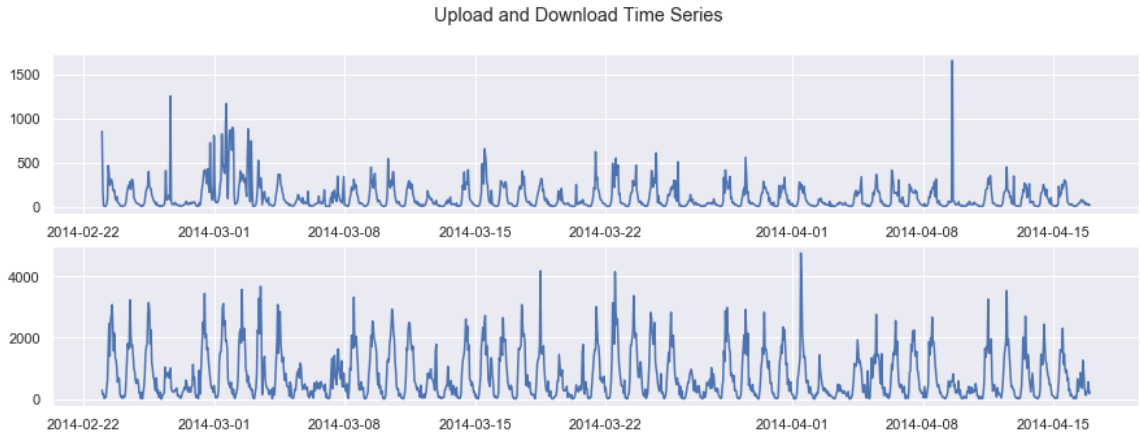
The authors in [29] perform time series analysis and prediction for LTE mobile traffic data. The algorithms used in this research are ARIMA and Exponential Smoothing. The data the analyze consists from 1352 cell stations across the city of Hong Kong, China. The datasets range starts from February 2014 and finished at March 2014, with a total of 21 days. Specifically, the dataset contains information for the Uplink and Downlink throughput, the timestamp; of an hourly frequency, and the cell ID with the according GPS coordinates. For evaluating their models, they use the RMSE and the R-squared metric. It is noted that the RMSE shows the difference between the prediction and real values, specifically the standard deviation of the residuals, where the R-squared shows the correlation between the prediction and actual value. Also, the maximum and optimal value for R-square is one (1). They proceed to explain that the stationary R-squared statistic is better for seasonal trends, therefore they chose this. Initially, they investigate the difference for downlink traffic between weekdays and weekends for the whole region (all cell stations). For modelling, they chose to separate their time series between weekdays and weekends, where they find the ARIMA model to be more accurate on weekdays, with a Stationary R-square value of 0.880 and the Exponential Smoothing model more accurate on weekends, with a Stationary R-square value of 0.730. Later, they use analysis on a single cells data, with RMSE as an evaluation metric, where Exponential Smoothing was found to be more accurate. The RMSE from this was 0.153.

4 Exploratory Data Analysis

In this Chapter, we will discuss the behavior the time series follows by using quantitative and qualitative techniques. With the term quantitative techniques, we evaluate statistical features that are expressed in arithmetic values. By the term qualitative techniques, we display insightful graphs of a time series, or alternative graphs generated from the time series. Also, graphs from sections generated in the quantitative section could be a part of the qualitative analysis. This section aims to explore features, trends and hidden knowledge that can be found in the data and be applied to achieve optimal forecasting.

Below we see the both time series plotted, with the Uplink data plotted at the top and the Downlink data at the bottom. The values in the y-axis are Megabits.

Figure 5



4.1 Dataset Description

The dataset is comprised of an excel spreadsheet with five (5) columns and one thousand two hundred and seventy-three (1273) rows. The two main attributes in our dataset

describe the Uplink and Downlink speed in Mbps for the 4G technology of a Mobile Network Operators (MNO) cell tower. The remaining three (3) columns describe the moment in time for each measurement. These are the Year, the Date and the Hour. The range in time of our dataset starts from 23rd February 2014 at midnight (00:00:00) and ends in the 16th of April 2014 at 23:00:00.

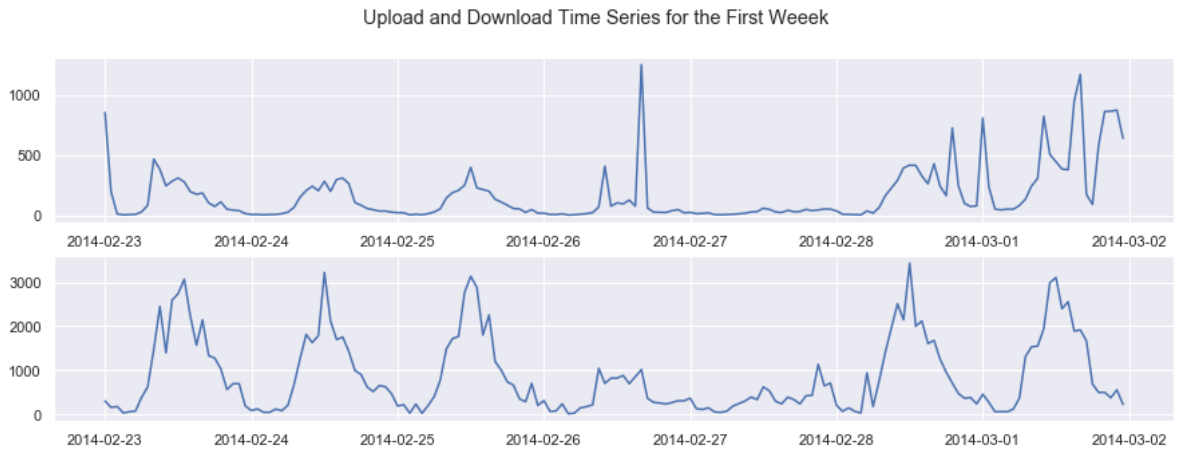
Therefore, the three-time related columns are combined into a single column in a Python pandas DataFrame with a datatype as a pandas timestamp. The format of the timestamp is:

$$YYYY - MM - DD\ HH{:}MM{:}SS$$

Also, the Uplink and Downlink data comprises a time series, so we use the timestamp as an index for the DataFrame. With the timestamp index, we can use efficient and insightful grouping and aggregating of the dataset for further analysis.

By looking at Figure 5, we observe a seasonal pattern on a weekly basis, with five (5) high peaks followed by two (2) lower peaks. This is more evident on the Download plot. However, we also note a second seasonal pattern on an hourly basis. To be more specific, we show a more detailed plot for the first week for both Upload and Download below in figure 6.

Figure 6



4.2 Missing or Erroneous Values

A time series is a sequence of data points that represent a quantity for every timestamp. These timestamps are successive and have equal space from one another in time. With real-life data, system failures and errors occur that obstruct taking measurements for every timestamp. Thus, missing or erroneous values may occur in a dataset.

The missing or erroneous values found within our dataset where either unknown (completely missing) or had a value of zero. Also, negative values are examined, where none were found.

Two missing rows of missing values were found in the data. They were for the same day, 7th of March 2014, for times 10:00 and 13:00. Both rows were missing Uplink and Downlink speeds. Hence, these four (4) values were replaced with the mean of the two (2) previous and following values, respectively.

4.3 Outlier Values

In Figure 5, we observe an outlying value on Wednesday 9th of April 2014 14:00:00 of 1655.73 Mbps for the upload series. The mean of the series however is 106.61 Mbps. Also, due to the dual seasonality of the time series, we further examine the values at hour 14:00:00 for day Wednesday with grouping and aggregating. The mean value for Wednesday at 14:00:00 is 273.42 Mbps, therefore this value is considered an outlier.

Also, in the Downlink time series again in Figure 5, we find another outlier at Monday 1st of April 2014 with a value of 4747.73 Mbps. The mean value for the Downlink data is 758.19 Mbps. By the same grouping and aggregation as previous, we find the mean for Mondays at 11:00:00 to be 1934.37 Mbps, therefore we consider this value an outlier.

4.4 Boxplots and Weekly Seasonal Pattern

In the 4th section of this chapter, boxplots of the dataset will be displayed, after grouping and aggregating the dataset by the weekday. The boxplots provide information regarding the median, the 25th and 75th percent quartile, the theoretical minimum and maximum and the outlying values.

The median is displayed with a straight line throughout the box and represents the value that separates the lower half from the higher half of the data within the distribution.

From the two quartiles, the interquartile range is observed, that is represented with the box.

The theoretical minimum and maximum values are calculated from the previously mentioned interquartile range. These are values that should be within 1.5 lengths of the interquartile range. These are represented by the “whiskers” and the final perpendicular line shows the minimum and maximum.

The remaining outlying values are the ones shown as dots, above or below the theoretical minimum and maximum respectively. From the boxplots, it is evident that a seasonal pattern is found, with least traffic on Wednesday and Thursday and more on the rest of the weekdays. Both boxplots have common starts and ends for the weekly seasonal pattern.

Figure 7

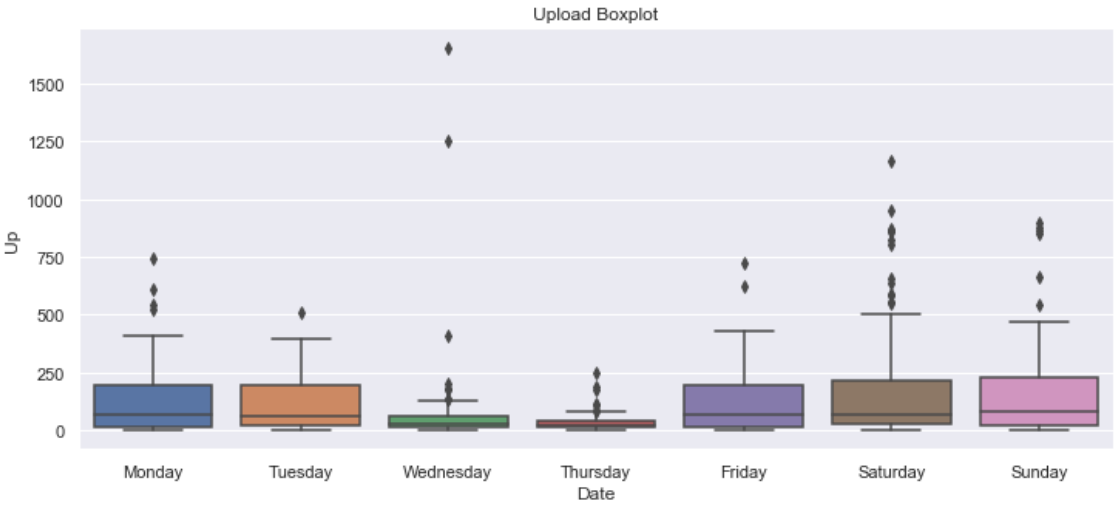
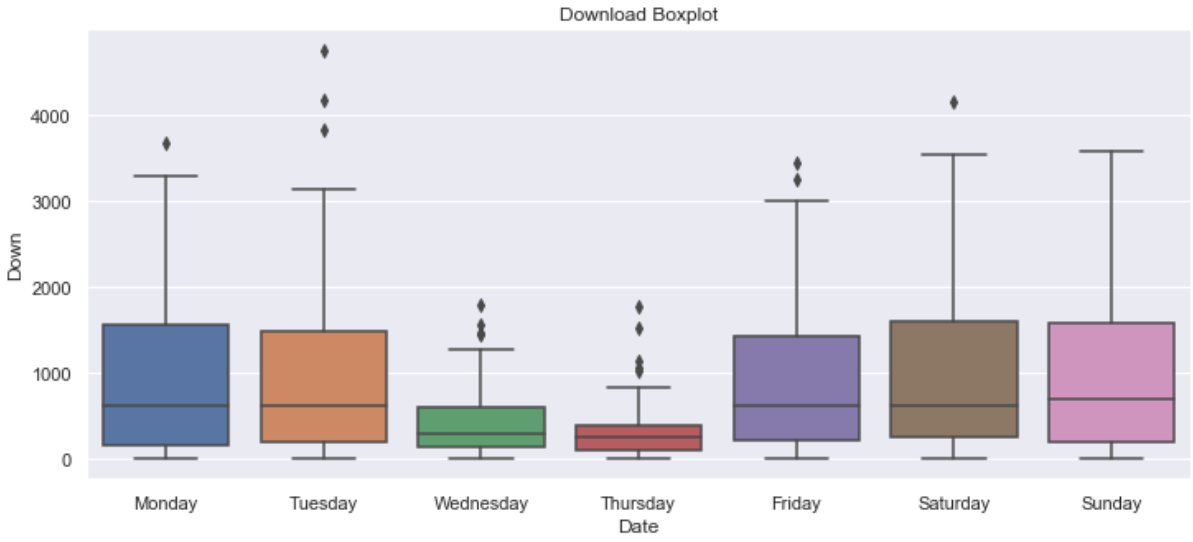


Figure 8



4.5 Time Series Decomposition

In Time Series Decomposition, the series is disaggregated by three main components. These components are the trend, the seasonal cycle and the residuals. The trend expresses a linear or non-linear dependency of the series throughout time. The seasonal component reveals patters that are repeated within the series and from the plot we can see the seasonal frequency and the seasonal components amplitude. The residuals are composed by the difference of the actual value from the trend and the seasonal component. Therefore, if the model is accurate, the residuals should have no correlation, pattern or trend. However, the previous is not always the case [30].

Figure 9

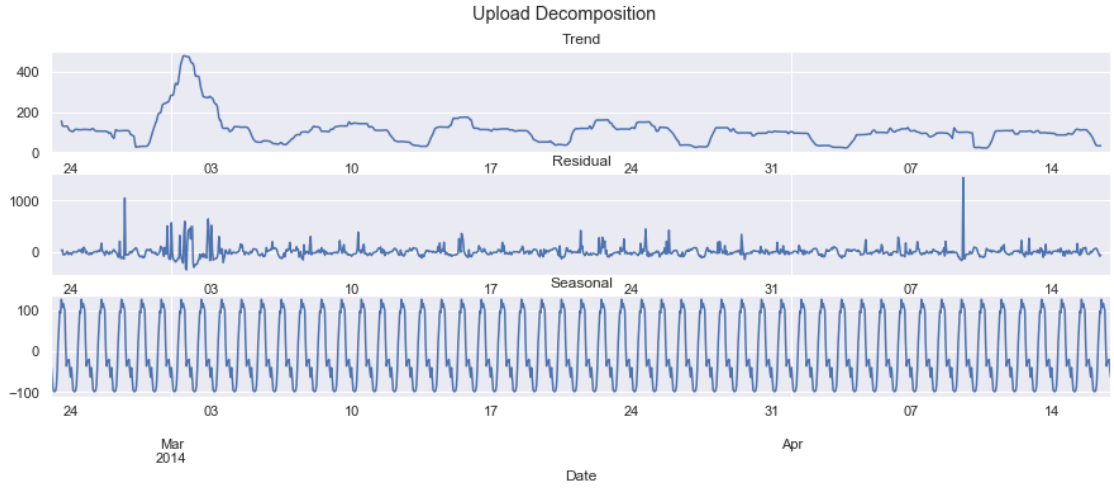
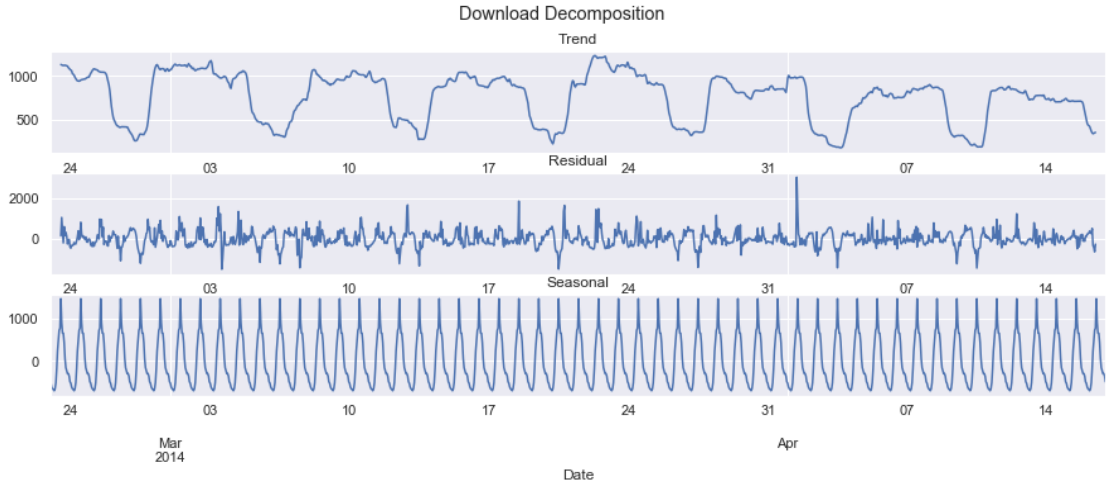


Figure 10



Above, two seasonal decompositions for upload and download are shown. The two seasonal patterns are clear and can be seen to be on an hourly frequency. The residuals do not show any pattern or autocorrelation for the timestamps. However, the trend components seem to have a seasonal pattern which is more evident in the Download series. As indicated in the previous section 3.4, a double seasonality within our dataset is confirmed.

4.6 Autocorrelation and Partial Autocorrelation Functions for SARIMA

From plotting the values of the Auto-Correlation Function (ACF) and Partial Auto-Correlation Function (PACF), candidate values for the orders of the S-Arima model can be discovered.

From the ACF plot a seasonal pattern is evident; therefore, a differencing order of 1 must be applied, first for the seasonal then for the non-seasonal orders.

Figure 11

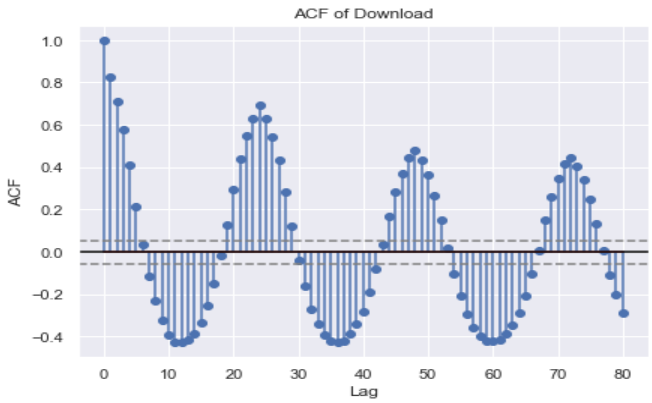
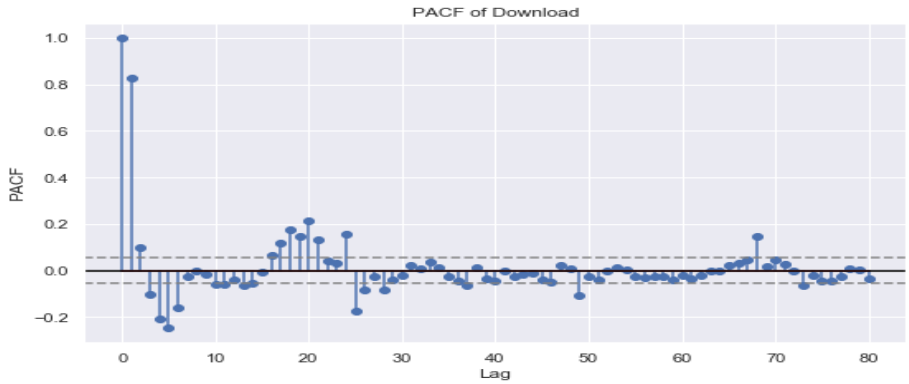


Figure 12



The next step is to repeat the plotting for the new differenced series. The PACF plot will define the order p, of the Auto Regressive model and the ACF plot will define orders for the Moving Average model. The seasonal orders are defined by significance of the lags close to the seasonal lag 24 and the non-seasonal orders for the first significant lags.

Equivalently, the same method for the upload time series are applied. These values for the order are indicative and will be applied into a grid search, although it is necessary to specify the maximum values for the orders.

5 Modelling

In the first step of creating, comparing and predicting forecasting models, the initial datasets are split into training and testing series. The training sets consists of 70% of the initial dataset, with calendar values from 23rd February 2014 00:00 up to 1st April 2014 01:00. After that, the test sets continue up to 16th April 2014 23:00.

5.1 Seasonal Arima

With the use of the python library pmdarima [31], a grid search is performed to find the optimal orders of the model. These candidate values were approximated with the use of Exploratory Data Analysis in Chapter 4. However, with the grid search, we were able to automatically compare each model by using the AIC criterion, and keep the one with the highest accuracy.

The parameters entered in the model were:

- Initial and maximum values for non-seasonal Auto Regressive order p
- Initial and maximum values for non-seasonal Moving Average order q
- Initial and maximum values for non-seasonal Differencing order d
- Initial and maximum values for seasonal Auto Regressive order P
- Initial and maximum values for seasonal Moving Average order Q
- Initial and maximum values for seasonal Differencing order D
- Initial and maximum values for non-seasonal Differencing order d
- Seasonal period m
- Stepwise was set to True

The stepwise parameter performs testing on various models with some parameters specified, e.g. p and d and keeps the model with the lowest AIC to grid search the remaining

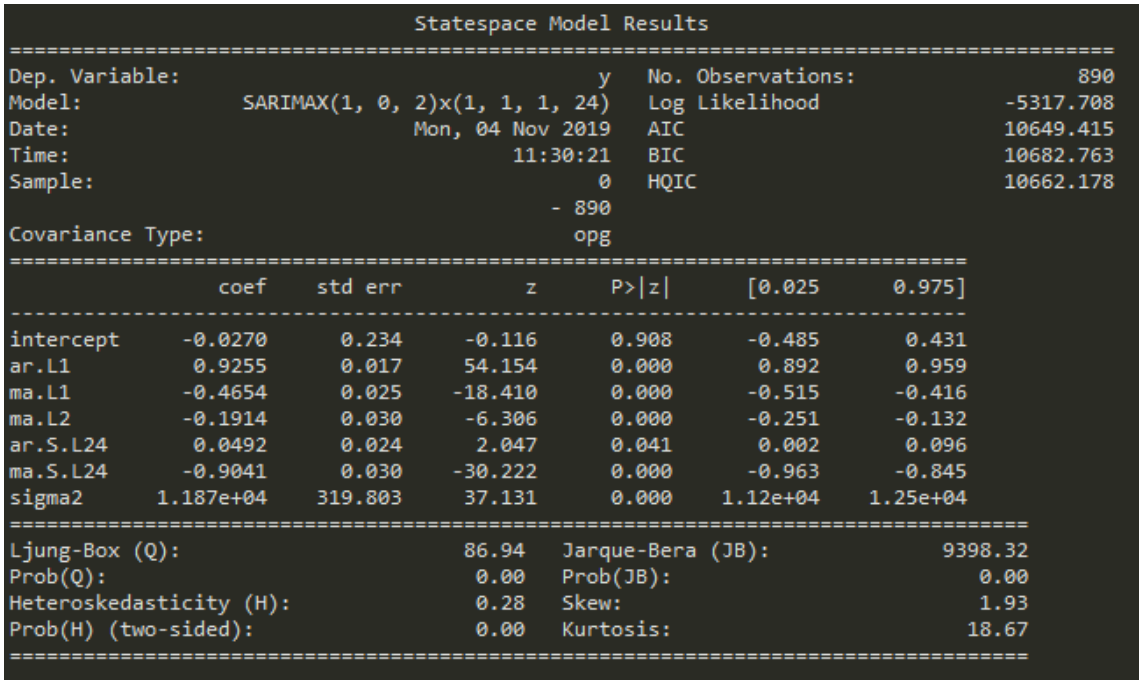
parameters. This method greatly boosts the speed of the grid search. Alternatively, with stepwise to be False, all the possible combinations of the model parameters are tested and then compared [31].

Having completed the grid search, we are given a model as a result. All the information regarding this model can be accessed with the .summary() command.

Below, both summaries are shown for upload and download time series.

First, is the summary for the model on the upload dataset.

Figure 13



Statespace Model Results

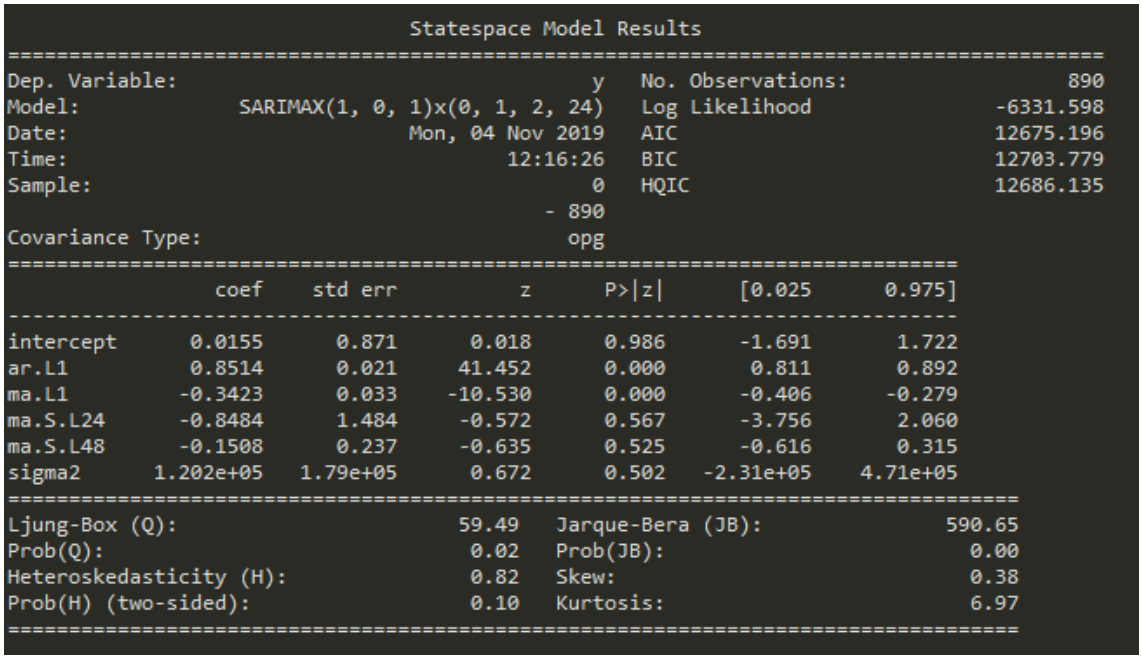
Dep. Variable:	y	No. Observations:	890
Model:	SARIMAX(1, 0, 2)x(1, 1, 1, 24)	Log Likelihood	-5317.708
Date:	Mon, 04 Nov 2019	AIC	10649.415
Time:	11:30:21	BIC	10682.763
Sample:	0	HQIC	10662.178
	- 890		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
intercept	-0.0270	0.234	-0.116	0.908	-0.485	0.431
ar.L1	0.9255	0.017	54.154	0.000	0.892	0.959
ma.L1	-0.4654	0.025	-18.410	0.000	-0.515	-0.416
ma.L2	-0.1914	0.030	-6.306	0.000	-0.251	-0.132
ar.S.L24	0.0492	0.024	2.047	0.041	0.002	0.096
ma.S.L24	-0.9041	0.030	-30.222	0.000	-0.963	-0.845
sigma2	1.187e+04	319.803	37.131	0.000	1.12e+04	1.25e+04

Ljung-Box (Q):	86.94	Jarque-Bera (JB):	9398.32
Prob(Q):	0.00	Prob(JB):	0.00
Heteroskedasticity (H):	0.28	Skew:	1.93
Prob(H) (two-sided):	0.00	Kurtosis:	18.67

Second is the summary for the model on the download dataset.

Figure 14



Statespace Model Results

Dep. Variable:	y	No. Observations:	890
Model:	SARIMAX(1, 0, 1)x(0, 1, 2, 24)	Log Likelihood	-6331.598
Date:	Mon, 04 Nov 2019	AIC	12675.196
Time:	12:16:26	BIC	12703.779
Sample:	0	HQIC	12686.135
	- 890		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
intercept	0.0155	0.871	0.018	0.986	-1.691	1.722
ar.L1	0.8514	0.021	41.452	0.000	0.811	0.892
ma.L1	-0.3423	0.033	-10.530	0.000	-0.406	-0.279
ma.S.L24	-0.8484	1.484	-0.572	0.567	-3.756	2.060
ma.S.L48	-0.1508	0.237	-0.635	0.525	-0.616	0.315
sigma2	1.202e+05	1.79e+05	0.672	0.502	-2.31e+05	4.71e+05

Ljung-Box (Q):	59.49	Jarque-Bera (JB):	590.65
Prob(Q):	0.02	Prob(JB):	0.00
Heteroskedasticity (H):	0.82	Skew:	0.38
Prob(H) (two-sided):	0.10	Kurtosis:	6.97

5.2 Exponential Smoothing

The first Exponential Smoothing algorithm was first proposed by [32] in 1957 and has since been developed. The exponential smoothing algorithm uses weighted averages of past observations while the weights decay exponentially the older the observation is. Therefore, recent observations have higher weights and higher impact on the predicted value. This method is generally used for time series with no specific trend or seasonality.

However, since we have time series with high seasonality, we investigate a version of Exponential Smoothing that can capture seasonal trends. This version is Holt-Winter's Exponential Smoothing algorithm that captures level, trend and a seasonal component. There are two versions of this algorithm, an additive model and a multiplicative model, equivalently to an ARIMA. Since multiplicative models are used for seasonal patterns that change proportionally to through time, which is not the case for our time series. This can be seen from the time series plots in Chapter 4 Exploratory Data Analysis [30].

The equations used to derive the forecast are displayed below:

$$\hat{y}_{t+h|t} = l_t + hb_t + s_{t+h-m(k+1)}$$

$$l_t = a(y_t - s_{t-m}) + (1+a)(l_{t-1} + b_{t-1})$$

$$b_t = \beta^*(l_t - l_{t-1}) + (1-\beta^*)b_{t-1}$$

$$s_t = \gamma(y_t - l_{t-1} - b_{t-1}) + (1-\gamma)s_{t-m}$$

The level shows the average between the seasonally adjusted observation and the non-seasonal. The trend shows the increase or decrease of the non-seasonal part through time.

In this algorithm, the only parameters to be specified by the user is an integer for a seasonal period, for which we used 24 for the hourly season. It should be noted that there was a runtime error when trying to use the weekly frequency.

5.3 TBATS

TBATS was proposed by the authors in [33] and combines Exponential Smoothing with Fourier terms for seasonality patterns. One of the main advantages of TBATS models is that the seasonal patterns may change over time, whereas in Seasonal ARIMA and Exponential Smoothing the same pattern is forced to repeat periodically, with respect to the period of the seasonal pattern [30]. However, the TBATS model is time consuming to run in comparison to the Holt-Winters Exponential Smoothing and to a Seasonal ARIMA forecasts.

In our work, the only parameters used in the TBATS estimator were the two seasonal periods found from Chapter 4 EDA, 24 for the hourly and 168 for the weekly seasons. Due to the multiple inserts regarding the seasonal data, we expect a higher accuracy of this model compared to the two previous.

6 Results

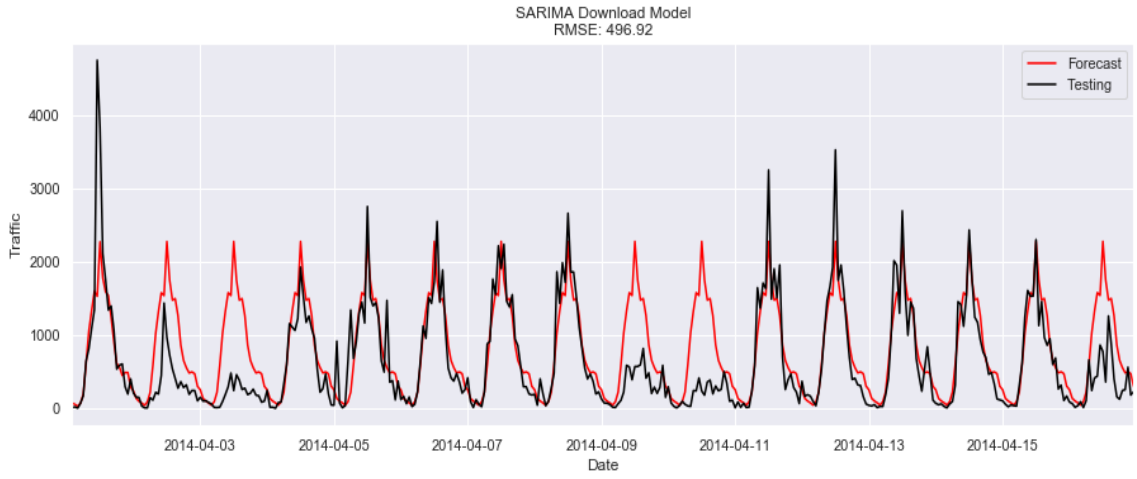
First, we start by showing the figures of the download forecast with the true values from the test set, for each algorithm. Then we proceed to repeat for the upload.

6.1 Download Dataset

6.1.1 Seasonal Arima

For the Seasonal Arima model, the optimal orders retrieved for (p, d, q) x (P, D, Q) were (1, 0, 1) x (0, 1, 2). It should be noted that a non-seasonal differencing doesn't exist, however a seasonal differencing of order 1 exists. Both these orders were expected, as there was no evident trend in the Seasonal Decomposition trend plots, but two seasonal ones were seen. Below the plot of the testing set and the SARIMA forecast is shown:

Figure 15



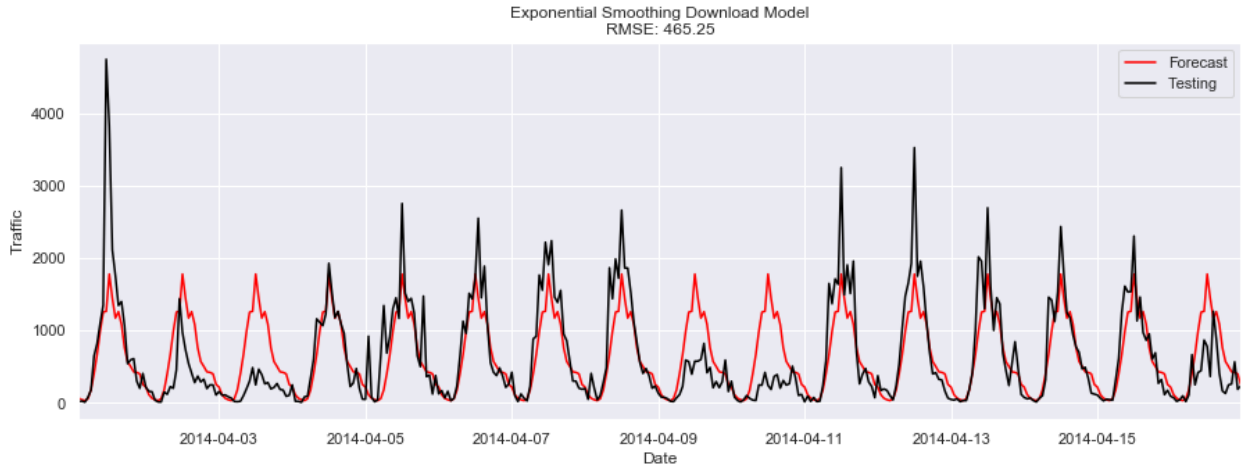
Also, a constant seasonal component is observed that is repeated with a daily period. This constant component is capturing accurately few days however is unable to provide good results for a daily season.

The minimum value of the seasonal period starts from tens of Megabits and reaches a maximum of about 2,200 Megabits.

6.1.2 Holt-Winters Exponential Smoothing

Exponential Smoothing was notably faster to apply in comparison to Seasonal Arima, since there a grid search was completed to find the appropriate orders. Here, only the seasonal period was inputted to the model and instantly training was completed. We timed this with python and the total training was done in 0.21seconds.

Figure 16

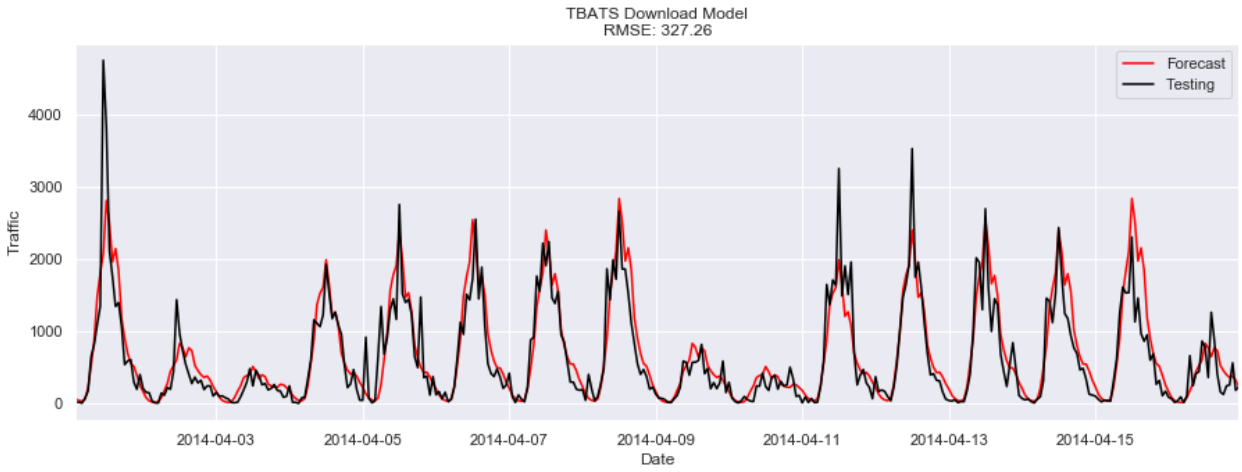


Here, better results were achieved with an RMSE around 6% lower than the SARIMA model, however the constant seasonal component yet again accurately approaches some days but fails to incorporate the weekly seasonal pattern. The seasonal low starts from tens of Megabits and reaches seasonal peaks of approximately 1,800 Megabits, which is significantly lower than the SARIMA model.

6.1.3 TBATS

As mentioned in Chapter Modelling, the TBATS algorithm can provide satisfying results for multiple and complex seasonalities. Here, multiple seasonal components can be seen for independent days throughout the week and the hours. It is evident that the TBATS model has provided the most accurate result with an RMSE of 327.26, almost 30% lower than Exponential Smoothing and approximately 34% lower than SARIMA. The time needed to train the TBATS model however was significantly higher than Exponential smoothing with a total of 132 seconds but still much faster than the grid search necessary to investigate the optimal model orders.

Figure 17



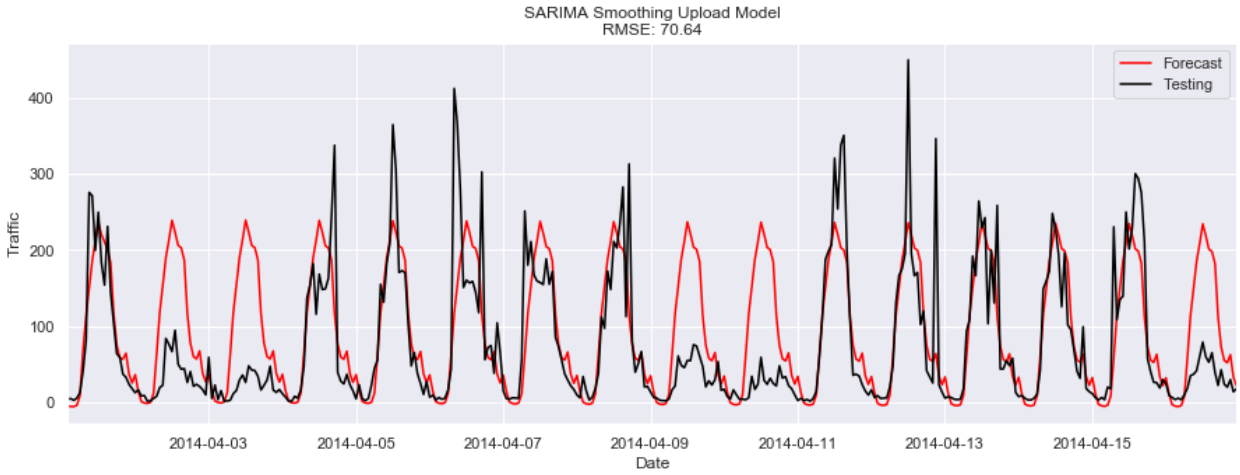
6.2 Upload Dataset

6.2.1 Seasonal Arima

The Seasonal Arima model provides again a constant seasonal component that is repeated through time. The higher values in the forecast reach up to 240 Mbps and start from tens of Mbps. Some irregularities are seen in the plot of the testing set, but these

are classified as an accidental pattern, that cannot be foreseen with a combination of the 2 seasonalities.

Figure 18



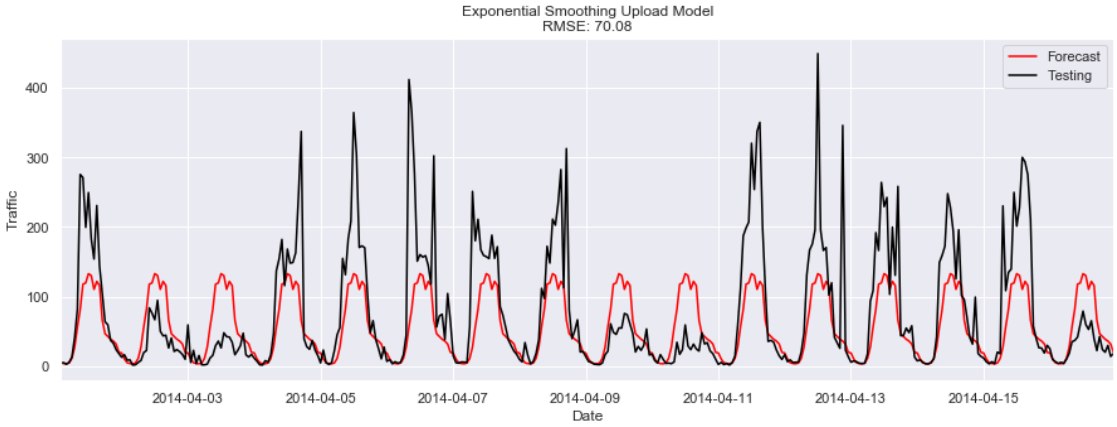
6.2.2 Holt-Winters Exponential Smoothing

Once again, the Exponential Smoothing algorithm provides a low RMSE. In the upload dataset the amplitude of the seasonal component is much smaller than the download dataset. Therefore, much lower scores in terms of RMSE were expected. Also, in proportion to their difference in mean values.

The total time for training and testing with Exponential smoothing was significantly faster, in accordance to the previous step with the download set, with 0.23 seconds. This tradeoff of computational time versus model accuracy could prove of high value when multiple applications of forecasting are necessary, in real time.

In terms of accuracy, the differences in the upload dataset between Exponential Smoothing and Seasonal Arima are negligible. In the download dataset a difference of up to 6% was seen, where here the difference is less than 1%. However, Exponential Smoothing has better results than SARIMA for days with minimal traffic and SARIMA for days with higher traffic.

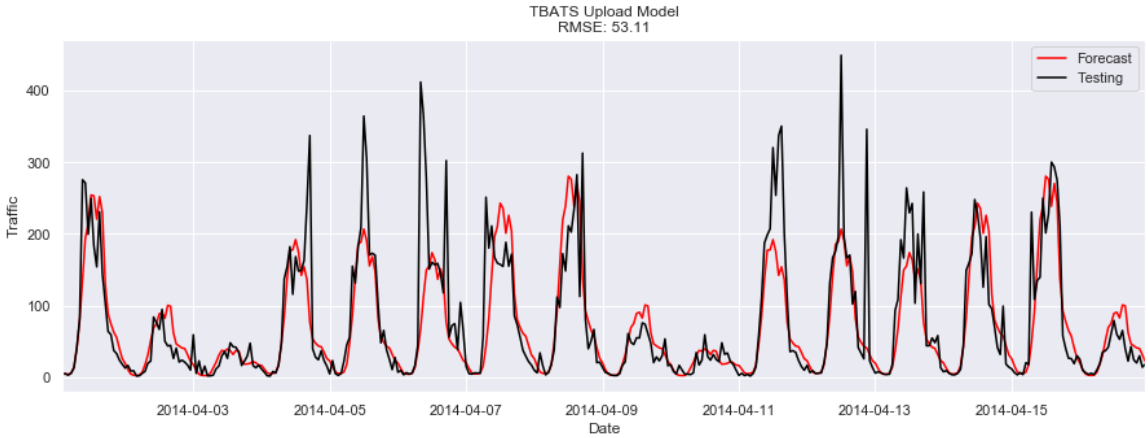
Figure 19



6.2.3 TBATS

The TBATS model provided the lower RMSE for both the Upload and the Download datasets. Here we have a RMSE of 53.11, which is 28% lower than both Exponential Smoothing and SARIMA models. The time required for training sums up to a total of 188 seconds. The accuracy in the weekly season shows close results to the actual for days with low traffic and reasonable approximations on the days with higher traffic. However, some abnormalities are evident, with some peaks being higher in one week than the other. Moreover, the hour these peaks took place is not consistent.

Figure 20



In the table below a summary for the resulting RMSE of the models is displayed for both upload and download.

Table 2

	Seasonal Arima	**Holt-Winters Exponential Smoothing**	**TBATS**
Upload	70.64	70.08	53.11
Download	496.92	465.25	327.26

7 Conclusions

In the final chapter we summarize the work, from data quality to the model theory up to the modelling procedure.

The purpose of this dissertation was to conduct a meticulous analysis of two time series of data traffic for a telecommunications cell tower. The main purpose during the modelling process was to maintain simplicity for the models in order to avoid overfitting and to perceive the characteristics of the time series.

7.1 Data

The dataset used in our work is real-life data with abnormalities. We are in position to break down the abnormalities and categorize them, namely the accidental abnormalities and the system failures. In more detail, system failures cause missing values or unknown values during the logging data process. These unknown values however can be approximated when investigating the neighboring values, given the number of successive missing values is very low. System failures can be caused from power outages, maintenance and upgrade of the system, etc. However, as accidental abnormalities, we describe the values that are measured correctly, but can be considered as outliers. These values can occur due to events taking place, such as football matches, concerts. Also, national holidays may increase use of voice and multimedia traffic on a telecommunication network. These characteristics can be seen during Exploratory Data Analysis, whether the higher than expected values occur in a short timeframe, or during the whole day.

In conclusion, not only acquiring correct data is important, but exploring within the patterns and understanding key points is of equal value.

7.2 Root Mean Square Error

The importance and reasoning of choosing the RMSE was not only found within the literature, but also because it is a good estimator for the standard deviation. Our model is based on the following equation:

$$\hat{y}_i - y_i = \varepsilon_i$$

Where y-hat is the observed value, y the predicted and ε the error. We assume the error to be a random variable with a mean μ of zero. If the mean is different from zero, then the model has a bias which should be considered within the predicted value. If we calculate the expected value from the RMSE we get:

$$E\left[\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}\right] = E\left[\frac{\sum_{i=1}^{n}\varepsilon_i^{\,2}}{n}\right] = \frac{1}{n}E\left[\sum_{i=1}^{n}\varepsilon_i\right] = E[\varepsilon] = Var(\varepsilon) + E[\varepsilon]^2$$
$$= \sigma^2 + \mu^2$$

However, μ is equal to zero therefore we get:

$$\sqrt{E\left[\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}\right]} = \sigma$$

7.3 Proposal for Future Works

Having performed forecasting for the time series, we note that accuracy was within reasonable limits. The process was completed by solely focusing on the time series, however there are other factors that could be considered. For future works, we propose incorporating external information relevant to mobile phone usage. Sources of external information can be in respect to weather conditions, where the authors in [34] find important effects of the microclimate as to where to position a cell tower. Another factor to be taken in consideration is spatial and regards the population within the towers range. To be more specific, in a metropolitan area, the population increases drastically during working hours and then decreases, whereas in suburbs, the opposite may occur.

As a last proposal, with completing the exploration and preparation of data for applying statistical methods, an extensive analysis was necessary. During this analysis, invaluable information was found with respect to the multiple seasonalities within the time series. Also, correlation found between lag features offered insights on how the time series should be processed. Therefore, we propose a future work on how these methods could be implemented in artificial neural networks.

Bibliography

[1] E. Oughton, Z. Frias, T. Russell, D. Sicker, and D. D. Cleevely, "Towards 5G: Scenario-based assessment of the future supply and demand for mobile telecommunications infrastructure," *Technol. Forecast. Soc. Change*, 2018.

[2] P. Gupta, "EVOLVEMENT OF MOBILE GENERATIONS : 1G T o 5G," *Int. J. Technol. Res. Eng.*, 2013.

[3] Cisco and S. Jose, "Cisco visual networking index (VNI) global mobile data traffic forecast update, 2017-2022 white paper," *Ca*, *Usa*, 2019.

[4] C. Zhang and P. Patras, "Long-term mobile traffic forecasting using deep Spatio-Temporal neural networks," in *Proceedings of the International Symposium on Mobile Ad Hoc Networking and Computing (MobiHoc)*, 2018.

[5] Q. K. U. D. Arshad, A. U. Kashif, and I. manzoor Qureshi, "A Review on the Evolution of the Cellular Technologies," *Acad. Perspect. Procedia*, 2019.

[6] P. Sharma, "Evolution of Mobile Wireless Communication Networks-1G to 5G as well as Future Prospective of Next Generation Communication Network," *Int. J. Comput. Sci. Mob. Comput. - not index*, 2013.

[7] C. S. Patil, R. R. Karhe, and M. A. Aher, "Review on Generations in Mobile Cellular Technology," *International Journal of Emerging Technology and Advanced Engineering*. 2012.

[8] S. Li, L. Da Xu, and S. Zhao, "5G Internet of Things: A survey," *Journal of Industrial Information Integration*. 2018.

[9] "Internet of Things forecast." [Online]. Available: https://www.ericsson.com/en/mobility-report/internet-of-things-forecast.

[10] IEEE, "IEEE 5G and Beyond Technology Roadmap White Paper," 2017.

[11] R. Talukdar and M. Saikia, "Evolution and Innovation in 5G Cellular Communication System and Beyond: A Study," 2014.

[12] CableFree, "LTE 4G & 5G Radio Access Network (RAN)." [Online]. Available: https://www.cablefree.net/wirelesstechnology/4glte/lte-4g-5g-radio-access-

network-ran/.

[13] Sd. Staff, “What Is the Radio Access Network?,” 2018. [Online]. Available: https://www.sdxcentral.com/5g/definitions/radio-access-network/.

[14] S. Louvros, K. Aggelis, and A. Baltagiannis, “LTE cell coverage planning algorithm optimising uplink user cell throughput,” in *Proceedings of the 11th International Conference on Telecommunications, ConTEL 2011*, 2011.

[15] “Top 10 Data Science Use Cases in Telecom.” [Online]. Available: https://activewizards.com/blog/top-10-data-science-use-cases-in-telecom/.

[16] R. A. Becker, C. Volinsky, and A. R. Wilks, “Fraud detection in telecommunications: History and lessons learned,” *Technometrics*, 2010.

[17] B. Huang, M. T. Kechadi, and B. Buckley, “Customer churn prediction in telecommunications,” *Expert Syst. Appl.*, 2012.

[18] C. M. Chen, “Use cases and challenges in telecom big data analytics,” *APSIPA Trans. Signal Inf. Process.*, 2016.

[19] V. Aggarwal, E. Halepovic, J. Pang, S. Venkataraman, and H. Yan, “Prometheus: Toward quality-of-experience estimation for mobile apps from passive network measurements,” in *Proceedings of the 15th Workshop on Mobile Computing Systems and Applications, HotMobile 2014*, 2014.

[20] “ARIMA lectures.” [Online]. Available: https://math.unm.edu/~ghuerta/tseries/week4_1.pdf.

[21] N. S. Nalawade and M. M. Pawar, “Forecasting telecommunications data with Autoregressive Integrated Moving Average models,” *2015 2nd Int. Conf. Recent Adv. Eng. Comput. Sci. RAECS 2015*, no. December, pp. 1–6, 2016.

[22] NIST/SEMATECH, “e-Handbook of Statistical Methods,” *E-handb. Stat. Methods*, 2012.

[23] P. Chen, A. Niu, D. Liu, W. Jiang, and B. Ma, “Time Series Forecasting of Temperatures using SARIMA: An Example from Nanjing,” *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 394, no. 5, 2018.

[24] R. E. Kass and A. E. Raftery, “Bayes factors,” *J. Am. Stat. Assoc.*, 1995.

[25] F. K. Oduro-gyimah and K. O. Boateng, “ANALYSIS AND MODELLING OF TELECOMMUNICATIONS NETWORK TRAFFIC : A TIME SERIES APPROACH,” no. June, 2018.

[26] S. Medhn, B. Seifu, A. Salem, and D. Hailemariam, “Mobile data traffic forecasting in UMTS networks based on SARIMA model: The case of Addis Ababa, Ethiopia,” in *2017 IEEE AFRICON*, 2017, pp. 285–290.

[27] G. Jia, P. Yu, P. Xiyuan, C. Qiang, Y. Jiang, and D. Yufeng, “Traffic forecasting for mobile networks with multiplicative seasonal ARIMA models,” in *ICEMI 2009 - Proceedings of 9th International Conference on Electronic Measurement and Instruments*, 2009.

[28] D. Tikunov and T. Nishimura, “Traffic prediction for mobile network using Holt-Winter’s exponential smoothing,” in *2007 15th International Conference on Software, Telecommunications and Computer Networks, SoftCOM 2007*, 2007.

[29] Xin Dong, Wentao Fan, and and Jun Gu, “Predicting LTE Throughput Using Traffic Time Series,” *ZTE Commun.*, 2015.

[30] R. J. Hyndman and G. Athanasopoulos, “Forecasting: Principles and Practice: Notes,” *OTexts*. 2014.

[31] “pmdarima documentation.” [Online]. Available: https://www.alkaline-ml.com/pmdarima/.

[32] C. C. Holt, “Forecasting trends and seasonals by exponentially weighted moving averages,” *ONR Memo.*, 1957.

[33] A. M. de Livera, R. J. Hyndman, and R. D. Snyder, “Forecasting time series with complex seasonal patterns using exponential smoothing,” *J. Am. Stat. Assoc.*, 2011.

[34] A. U. Usman, O. U. Okereke, and E. E. Omizegba, “Instantaneous GSM Signal Strength Variation with Weather and Environmental Factors American Journal of Engineering Research (AJER),” *Am. J. Eng. Res.*, no. 3, pp. 104–115, 2015.